CS 726: Homework #2

Posted: 02/11/2020, due: 02/24/2020 by 5pm on Canvas

Please typeset or write your solutions neatly! If we cannot read it, we cannot grade it.

Note: You can use the results we have proved in class – no need to prove them again.

Q 1. Recall the Gauss-Southwell rule for basic descent methods that we saw in class: $\mathbf{d}_k = -\nabla_{i_k} f(\mathbf{x}_k)\mathbf{e}_{i_k}$, where $i_k = \mathrm{argmax}_{1\leq i\leq n} |\nabla_i f(\mathbf{x}_k)|$ and $\mathbf{e}_{i_k}$ is the vector that has 0 in all coordinates except for i_k, where it equals 1 (it is the $i_k{}^{\text{th}}$ standard basis vector). Same as in the class, we assume that f is L-smooth. Prove that there exists $\alpha > 0$ such that the Gauss-Southwell rule applied for an appropriate step size α_k satisfies:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

How would you choose α_k? What can you say about the convergence of this method (discuss all three cases we have covered in class: nonconvex and bounded below, convex, strongly convex)? [10pts]

Q 2. Exercise 8 from Chapter 3 in Recht-Wright. Notation from the exercise: $\mathbf{x}_\star = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. You don't need to worry about getting the same constants as stated there, being off by constant factors (up to 4) is fine. [15pts]

Q 3 (Bregman Divergence)**.** Bregman divergence of a continuously differentiable function $\psi : \mathbb{R}^n \to \mathbb{R}$ is a function of two points defined by

$$D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle .$$

Equivalently, you can view Bregman divergence as the error in the first-order approximation of a function:

$$\psi(\mathbf{x}) = \psi(\mathbf{y}) + \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + D_\psi(\mathbf{x}, \mathbf{y}).$$

(i) What is the Bregman divergence of a simple quadratic function $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$, where $\mathbf{x}_0 \in \mathbb{R}^n$ is a given point? [5pts]

(ii) Given $\mathbf{x}_0 \in \mathbb{R}^n$ and some continuously differentiable $\psi : \mathbb{R}^n - \mathbb{R}$, what is the Bregman divergence of function $\phi(\mathbf{x}) = \psi(\mathbf{x}) + \langle \mathbf{x}_0, \mathbf{x}\rangle$? [5pts]

(iii) Use Part (ii) and the definition of Bregamn divergence to prove the following 3-point identity:

$$(\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n): \quad D_\psi(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{z}, \mathbf{y}) + \langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z}\rangle + D_\psi(\mathbf{x}, \mathbf{z}). \quad \text{[5pts]}$$

(iv) Suppose I give you the following function: $m_k(\mathbf{x}) = \sum_{i=0}^k a_i\psi_i(\mathbf{x})$, where $\psi_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{x}_i\|_2^2 + \langle \mathbf{b}_i, \mathbf{x} - \mathbf{x}_i\rangle$, where $\{a_i\}_{i\geq 0}$ is a sequence of positive reals and $\{\mathbf{b}_i\}_{i=0}^k$, $\{\mathbf{x}_i\}_{i=0}^k$ are fixed vectors from $\mathbb{R}^n$. Define $\mathbf{v}_k = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^n} m_k(\mathbf{x})$ and $A_k = \sum_{i=0}^k a_i$. Using what you have proved so far, prove the following inequality:

$$(\forall \mathbf{x} \in \mathbb{R}^n): \quad m_{k+1}(\mathbf{x}) \geq m_k(\mathbf{v}_k) + a_{k+1}\psi_{k+1}(\mathbf{x}) + \frac{A_k}{2}\|\mathbf{x} - \mathbf{v}_k\|_2^2. \quad \text{[5pts]}$$

Q 4. In class, we have analyzed the following variant of Nesterov's method for L-smooth convex optimization:

$$\mathbf{x}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1}$$
$$\mathbf{v}_k = \mathbf{v}_{k-1} - a_k\nabla f(\mathbf{x}_k)/L$$
$$\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k),$$

where L is the smoothness constant of f, $a_0 = A_0 = 1$, $\frac{{a_k}^2}{A_k} = 1$, $A_k = \sum_{i=0}^{k} a_i$. We take $\mathbf{x}_0 \in \mathbb{R}^n$ to be an arbitrary initial point and $\mathbf{y}_0 = \mathbf{v}_0 = \mathbf{x}_0 - \nabla f(\mathbf{x}_0)/L$.

Prove that we can state Nesterov's method in the following equivalent form:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{y}_{k-1} + \frac{a_k}{A_k}\Big(\frac{A_{k-1}}{a_{k-1}} - 1\Big)(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}), \\ \mathbf{y}_k &= \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k). \end{aligned} \tag{1}$$

Hint: It is helpful to first prove that $\mathbf{y}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k$. [10pts]

Q 5 (Coding Assignment)**.** In the coding assignment, we will compare different optimization methods discussed in class on the following problem instance: $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$, where $f(\mathbf{x}) = \frac{1}{2}\langle \mathbf{M}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$, $\mathbf{b}$ is a vector whose first coordinate is $1 - \frac{1}{n}$ while the remaining coordinates are $\frac{1}{n}$, and $\mathbf{M}$ is the same matrix we saw in Q 8 of Homework #1. We will take the dimension to be $n = 200$. Matrix $\mathbf{M}$ and vector $\mathbf{b}$ can be generated using the following Matlab code:

```
k = n;
M = diag(2*[ones(k, 1); zeros(n-k, 1)], 0)...
    + diag([-ones(k-1, 1); zeros(n-k, 1)], -1)...
    + diag([-ones(k-1, 1); zeros(n-k, 1)], 1);
M(n,1) = - 1;
M(1,n) = -1;
b = -1/n * ones(n, 1);
b(1) = b(1) + 1;
```

Observe that you can compute the minimizer $\mathbf{x}^*$ of f given $\mathbf{M}$ and $\mathbf{b}$, and thus you can also compute $f(\mathbf{x}^*)$. It is possible to show that the top eigenvalue of $\mathbf{M}$ is $L = 4$.

Implement the following algorithms:

1. Steepest descent with the constant step size $\alpha_k = 1/L$.

2. Steepest descent with the exact line search.

3. Lagged steepest descent, defined as follows: Let α_k be the exact line search steepest descent step size corresponding to the point $\mathbf{x}_k$. Lagged steepest descent updates the iterates as: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_{k-1}\nabla f(\mathbf{x}_k)$ (i.e., the step size "lags" by one iteration).

4. Nesterov's method for smooth convex minimization.

Initialize all algorithms at $\mathbf{x}_0 = \mathbf{0}$. All your plots should be showing the optimality gap $f(\mathbf{x}) - f(\mathbf{x}^*)$ (with $\mathbf{x} = \mathbf{y}_k$ for Nesterov and $\mathbf{x} = \mathbf{x}_k$ for all other methods) on the y-axis and the iteration count on the x-axis. The y-axis should be shown on a logarithmic scale (use `set(gca, 'YScale', 'log')` after the figure command in Matlab).

(i) Use a single plot to compare steepest descent with constant step size, steepest descent with the exact line search, and Nesterov's algorithm. Use different colors for different algorithms and show a legend with descriptive labels (e.g., SD:constant, SD:exact, and Nesterov). Discuss the results. Do you see what you expect from the analysis we saw in class?

(ii) Use a single plot to compare Nesterov's algorithm to lagged steepest descent. You should, again, use different colors and a legend. What can you say about lagged steepest descent? How does it compare to Nesterov's algorithm?

(iii) Modify the output of Nesterov's algorithm and lagged steepest descent: you should still run the same algorithms, but now your plot at each iteration k should show the lowest function value up to iteration k for each of the two algorithms. Discuss the results.

You should turn in both the code (as a text file) and a pdf with the figures produced by your code together with the appropriate answers to the above questions. [45pts]

Note: Your code needs to compile without any errors to receive any points for the coding assignment.